Region One
Vegetation Classification, Mapping, Inventory and Analysis Report

Numbered Report 12-36-v2 April, 2012

Flathead National Forest VMap Accuracy Assessment; Version 12

Stephen R. Brown Jr.

USDA Forest Service, Region 1
Engineering Staff
200 E. Broadway, Missoula, MT 59802

1. Accuracy Assessment Defined

Accuracy assessment is an essential part of any remote sensing project. It provides the basis of comparison for different methods and/or sensors. It provides information regarding the reliability and usefulness of remote sensing techniques for a particular application. Most importantly accuracy assessment provides a validation of the data, giving an indication of reliability of the classification, so that managers are fully informed throughout the decision making process. Too often vegetation and other maps are used without a clear understanding of their reliability. A false sense of security about the accuracy of the map may result in an inappropriate use of the map and important decisions may be made based on data with unknown and/or unreliable accuracy. Estimates of overall map accuracy and confidence of individual map classes can be inferred from an error matrix derived from the comparison of known reference sites to mapped data. Although quantitative accuracy assessment can be time-consuming and expensive, it is an integral part of any vegetation-mapping project.

Accuracy, however, is not a state variable. It is very important to evaluate the results of any accuracy assessment in the context of the intended analysis application and the management decision the data and analyses are intended to support. This evaluation needs to balance the desired level of precision (i.e., the level of thematic detail) with the desired level of accuracy (i.e., spatial location of a given attribute). For many analyses, detailed thematic classes are aggregated to produce more generalized classes that typically increases the accuracy of a given map. It is appropriate in these instances to assess the accuracy of the aggregated classes rather than characterize the aggregations with the detailed assessment. It may even be appropriate to aggregate some classes based on the structure of the error, provided that the aggregations meet the analysis objectives. It is also important to determine the level of uncertainty that is acceptable to support a particular management decision.

Quantitative accuracy assessment depends upon the collection of reference data with which to compare the map product in question. It is therefore assumed that the reference data is “truth”, that is 100% correct. Reference data can be obtained via field site visits, photo-interpretation, existing plot data, or a combination of these methods. Statistical validity of the sample, however, is most easily maintained through a random selection of sites which can make the acquisition of reference data both cost and time prohibitive. To overcome this limitation a method has been devised that incorporates a random sample selection with field site visits, photo-interpretation, and existing plot data through the use of aerial resource photography and Forest Inventory and Analysis (FIA) plot data. Forest Inventory and Analysis data have been collected in a standardized, grid-like fashion across the United States for approximately 70 years. Data collected by FIA contain information about forest characteristics such as species composition, size-class, canopy coverage, health, and growth rates to name just a few. Having been collected in a consistent manner and distributed across the landscape as a network of points the information recorded by the FIA program provides a base from which an independent, systematic, assessment of VMap class accuracy can be conducted. The FIA data is not, however, collected for mapping purposes and is not directly comparable to the VMap product. It can, though, be intersected with VMap polygons to produce the random sample selection needed and then be used to inform an analyst as to the general composition of the stand in question and guide them in the photo-interpretation process.

After completion of the photo-interpretation process for all FIA intersected polygons, comparisons of these data to the mapped elements are then tabulated and presented in an error matrix, where the rows represent values of the map, and columns represent values of the reference data. Tabulated values across the diagonal of the matrix describe the number of times map and reference data sites have equal values. Conversely, the off-diagonal table elements quantify errors of either inclusion or exclusion of particular classes. Errors of inclusion are shown on the horizontal axis of classes, while errors of exclusion are shown on the vertical axis. Large numbers of inclusion or exclusion between two or more classes indicate a high degree of inter-class confusion and generally indicate a lower quality map. To illustrate these concepts, an error matrix quantifying the level of agreement in a theoretical lifeform map is given below as Table 1.

Table 1. *Error matrix of a theoretical lifeform map, with overall map accuracy of 74%*

		Reference Data Classes				
		Forest	*Shrub*	*Herbaceous*	*Water*	**Map Total**
Map Data Classes	*Forest*	**65**	4	22	24	115
	Shrub	6	**81**	5	8	100
	Herbaceous	0	11	**85**	19	115
	Water	4	7	3	**90**	104
	Ref. Total	75	103	115	141	434

Once an error matrix table has been created, several useful measures of map accuracy can be computed, including overall, producer, and user metrics. Overall accuracy is a common metric that describes how well the map compares to a reference dataset as a whole. Producer accuracy focuses on errors of exclusion and thus is a term that describes the number of samples that were incorrectly classed. User accuracy, on the other hand, is based on errors of inclusion and therefore reflects the probability that a feature of the map actually represents that category on the ground. Regardless of the measurement used, the robustness of the metric is largely dependent on the number of samples that were used for comparison. In the best case scenario a similar number of samples will be available for each map class, and each class will have a large number of samples, which generally means more than 30 instances. It is unfortunate, but an assessment of individual class accuracy cannot be conducted when there are an insufficient number of reference samples available. In such cases users of the map should be aware that while the error in some map classes is not quantified in an error matrix, it can be assessed either through additional reference data collection, or via systematic field review of the classification.

Overall Accuracy is computed by dividing the total number of correct samples by the total number of assessment sites found in the bottom right cell of the error matrix table.
It is often the most commonly reported accuracy measure because it takes advantage of samples

from all classes. Not all map classes will have large enough samples available for comparison. With Table 1 as an example, it can be seen that 434 sites were evaluated against their known condition on the ground. By adding the total number of times mapped classes were in agreement with their known condition and dividing that total by the total number of sites that were evaluated the overall accuracy of the map can be assessed as follows:

[*Forest* (**65**) + *Shrub* (**81**) + *Herbaceous* (**85**) + *Water* (**90**) = 321] / 434 = **74%**

Producer Accuracy is the probability of a reference site being correctly classified, and is calculated by dividing the total number of correctly mapped sites for a class by the total number of reference sites for that class. Using data from Table 1, Producer's class accuracy values are assessed as follows in Table 2:

Table 2. *Computation of Producer Map Accuracy*

Map Class	# of correct sites	# of all reference sites	Relative Accuracy (%)
Forest	**65** divided by	**75**	**= 87**
Shrub	**81** divided by	**103**	**= 79**
Herbs	**85** divided by	**115**	**= 74**
Water	**90** divided by	**141**	**= 64**

User Accuracy is the probability that a feature on the map actually represents that category on the ground, and is calculated by dividing the number of agreements for a category by the total number of sites that were mapped into that category. Using data from Table 1, User class accuracy values are assessed as follows in Table 3:

Table 3. *Computation of User Map Accuracy*

Map Class	# of correct sites	# of all mapped sites	Relative Accuracy (%)
Forest	**65** divided by	**115**	**= 57**
Shrub	**81** divided by	**100**	**= 81**
Herbs	**85** divided by	**115**	**= 74**
Water	**90** divided by	**115**	**= 87**

For a more detailed description of the accuracy assessment process used to complete the eastside R1 VMap accuracy assessment see document 'R1-VMap Accuracy Assessment Procedures for Region 1', Vanderzanden et al, 2009. CMIA # 09-11.

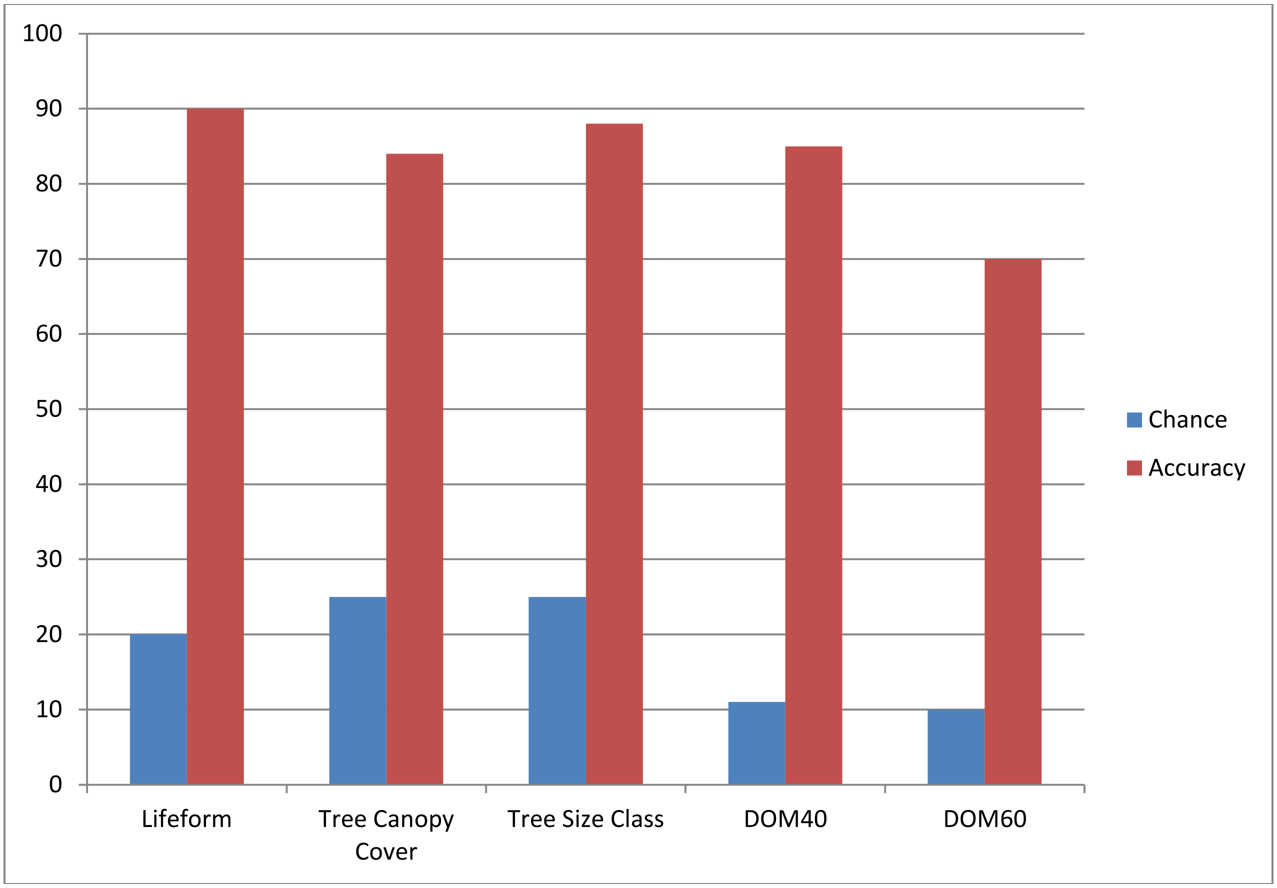
2. Results and Discussion

For the Flathead National Forest (FNF) VMap accuracy assessment, there were a total of 384 FIA plots available for assessment. Of these, 280 were suitable for comparison for Tree Canopy Cover, 260 for Tree Size Class, 259 for DOM MID 40, and 259 for DOM MID 60. Disturbance plot location (i.e., located within a deep shadow of a cliff) rendered the remainder of the plots unusable. Overall the resulting map products created for the landscapes encompassed by the FNF VMap show exceptional accuracies.

In each of the forested analysis areas, error matrices have been constructed for the mid level dominance plurality classes (DOM_MID_60, DOM_MID_40), four classes of tree canopy cover (10-24.9%, 25-39.9%, 40-59.9%, 60+%), and four classes of tree size (0-4.9", 5-9.9", 10-14.9”, = 15+".) A separate analysis for the Lifeform class was conducted based on a stratified random sample size of 1050 samples.

Overall Accuracy is a measure of the agreement between the sampled sites and mapped classes corresponding to those sites. It is simply the sum of the number of sites that agree divided by the total number of sites that were compared. As such, Overall Accuracy says nothing about individual class accuracy; rather it provides the interpreter with a measure of classification quality as a whole. It is important to consider that the value of this measure is influenced by the number of comparisons that are made in each of the classes. This can be overcome by either making the sample size the same for each class or by normalizing the elements of the error matrix. To be meaningful, each class being compared would have at least thirty samples. When such criteria are not met the assessment of classes with small sample sizes is not very meaningful, or realistic, and the Overall Accuracy statistic is the only remaining measure of map accuracy with any reliability.

Oftentimes Users do not have realistic expectations of what an acceptable level of map accuracy should be. Accuracy is generally evaluated based one’s inherent familiarity with the academic system of grading. This is a flawed comparison, however, as map accuracy is not a static variable but changes in meaning with both application and the number of classes that are being represented. A more useful interpretation of map accuracy, then, would be a comparison to the probability of chance agreement between classes. For example, a map of with 12 classes with an accuracy of 65% seems fairly limited based on an academic scale, and seems only somewhat better than flipping a coin (which has a 50% chance of getting the right answer). However when that is compared to chance agreement (which would be ~8% with 12 classes) it is seen that a map with 65% accuracy is actually 8 times better than chance and provides the User with a high degree of confidence in the placement of classes across the map. Graph 1, below, shows a comparison of the Overall Accuracy for each VMap product to the probability of chance agreement based on the number of classes. A full discussion of the individual product accuracies follows.



Graph 1. Percent Overall Accuracy versus Chance Agreement by VMap Product.

2a. Lifeform Accuracy Assessment

While the FIA plot design was the basis for site selection with the other classes, this proved to be insufficient for a proper Lifeform assessment. Therefore, following the recommendations of Stehman and Czaplewski (1998), a stratified random sample was designed for the Lifeform class, with 1050 samples collected overall, divided into 210 samples per class.

It can be seen in Table 4 below that the Overall accuracy estimate for the Lifeform Class of the FNF VMap is 90%. This is exceptional accuracy given the size and complexity of the FNF landscape. The matrix does show, however, that the majority of the confusion lies between the grass and sparsely-vegetated lifeforms, with grass perhaps being over estimated in the upper elevations. It is difficult to ascertain this with a high degree of certainty though as the samples were collected and evaluated remotely and reliably distinguishing between greater than 10% grass cover and less than 10% grass cover in the upper elevations can be difficult. Another area of confusion is between that of grass and shrub where, again, it can be difficult to distinguish between low covers of each remotely. Considering that the vast majority of analysis will deal with the Tree class, this area of confusion should not raise any serious concerns. The Tree class is shown to be correctly mapped 97% of the time, which gives the User confidence that involves trees will be confined to Tree types.

Table 4. Lifeform Class Error Matrix

Flathead National Forest VMap V12; Lifeform							
	AA Data						
VMap Class	**Grass**	**Shrub**	**Tree**	**Water**	**Sparse-Veg**	**Grand Total**	**User Accuracy**
Grass	**173**	4	6		27	**210**	**82%**
Shrub	14	**177**	8	5	6	**210**	**84%**
Tree		4	**204**	2		**210**	**97%**
Water	1		1	**202**	6	**210**	**96%**
Sparse-Veg	12	1	8	1	**188**	**210**	**90%**
Grand Total	**200**	**186**	**227**	**210**	**227**	**1050**	**Overall Accuracy**
							90%

2b. Dominance 60% Plurality (DOM_MID_60) Accuracy Assessment

Based on the numbers seen in Table 5, below, it can be seen that the Overall map accuracy for DOM MID 60 is 70%. Unfortunately, nothing can be said for the individual map classes of PIPO, LAOC, PIEN, POPUL, and HMIX as there is an insufficient number of samples present to make a statistically valid estimate. It general recommended to have a minimum N of 20-30 samples for significance. Of the remaining classes, the three most abundant are PSME, PICO, and ABLA with the least amount of error being in the ABLA class. The most error is between the IMIX/TMIX classes the "pure" types. This is common and is somewhat attributable to the R1 Vegetation Classification system that is used for mapping purposes. The problem being that these IMIX/TMIX classes look like everything and it can be hard to consistently distinguish between them and the "pure" types. Overall, however, the "pure" types are well represented and any error in their location is not significant enough to adversely impact analysis.

Table 5. DOM_MID_60 Class Error Matrix

Flathead National Forest VMap V12; DOM MID 60												
	AA Data											
VMap Class	PIPO	PSME	LAOC	PICO	ABLA	PIEN	POPUL	IMIX	TMIX	HMIX	Grand Total	User Accuracy
PIPO	1										1	N/A
PSME		36	1	1	3			5	7		53	68%
LAOC		2	8	1	1			2	3		17	N/A
PICO		4		39	5	2		3	2		55	71%
ABLA				1	45	3		1	4	1	55	82%
PIEN			1		1	12		1	1		16	N/A
POPUL							1				1	N/A
IMIX				1				11	3		15	N/A
TMIX		1			13	1		1	24		40	60%
HMIX										1	1	N/A
Grand Total	1	43	10	43	68	18	1	24	44	2	254	Overall Accuracy
Producer Accuracy	N/A	84%	N/A	91%	66%	N/A	N/A	46%	55%	N/A		70%

2c. Dominance 40% Plurality (DOM_MID_40) Accuracy Assessment

There is a slight improvement in the accuracy assessment for DOM_MID_40 over DOM_MID_60 (Table 6), going from 70% to 85%. Again, this is due in part to the classification definition of the class allowing for the inclusion of other species within each mixed class label rather than the straight IMIX/TMIX label. Due to the limitations of the FIA sample design we are again limited in what can be said of the individual classes outside of those 5 with sufficient N. With the "Big 3" – PSME, PICO, and ABLA – from the discussion above there is a marked improvement for the MX-ABLA and MX_PSME classes, a gain of > 10%, while the MX-PICO exhibits little change. One difference is that there are now enough samples to be able to start to say something about MX-LAOC and MX-PIEN, both of which show very high User accuracy. Also, it can be seen in the error matrix that there is zero confusion between the coniferous types and MX-POPUL, that there are no instances where something else is mislabeled as MX-POPUL. This indicates that where an object is labeled as MX-POPUL one can be fairly certain it will be that and nothing else.

Table 6. DOM_MID_40 Class Error Matrix

Flathead National Forest VMap V12; DOM MID 40											
	AA Data										
VMap Class	MX-PIPO	MX-PSME	MX-LAOC	MX-PICO	MX-ABLA	MX-PIEN	MX-PIAL	MX-POPUL	HMIX	Grand Total	User Accuracy
MX-PIPO	1									1	N/A
MX-PSME		45		2	5	4				56	80%
MX-LAOC		2	23	1	1					27	85%
MX-PICO		5		43	6	2	1			57	75%
MX-ABLA					82	2	4		1	89	92%
MX-PIEN		1	1			20				22	91%
MX-PIAL							4			4	N/A
MX-POPUL								1		1	N/A
HMIX									2	2	N/A
Grand Total	1	53	24	46	94	28	9	1	3	259	Overall Accuracy
Producer Accuracy	N/A	85%	96%	93%	87%	71%	N/A	N/A	N/A		85%

2d. Tree Canopy Cover Accuracy Assessment

The tree canopy cover error matrix (Table 7) shows an error distribution that is fairly typical of categorized variables, with most of the confusion existing between the adjacent classes. This is not surprising given that field data collection protocols only require that accuracies be within plus or minus one class. Also, a review of the FIA plot data reveals that much of the forested area is right on the edge of a class, rarely at the midpoint. For example, many stands show a canopy cover estimate of 42%, which is just inside the 40-59.9% class but may be easily confused with the upper end of the 25-39.9% class. All in all, though, the tree canopy cover map product performs very well. It is seen from the table that overall the estimates for each class are close, with the largest degree of confusion between the 40-60% and 60%+ Canopy Cover classes, with an underestimate of the more dense canopy class.

Table 7. Tree Canopy Cover Class Error Matrix

Flathead National Forest VMap V12; Tree Canopy Cover						
	AA Data					
VMap Class	10-25% Cover	25-40% Cover	40-60% Cover	60%+ Cover	Grand Total	User Accuracy
10-25% Cover	16	1	3		20	80%
25-40% Cover		31	9	3	43	72%
40-60% Cover	1	1	71	27	100	71%
60%+ Cover			1	116	117	99%
Grand Total	17	33	84	146	280	Overall Accuracy
Producer Accuracy	94%	94%	85%	79%		84%

2e. Tree Size Class Accuracy Assessment

Once again, against the common expectation, tree size is the top performer of the VMap classes (Table 8). Presumably this can be attributed to two things. The first being the inclusion of the NAIP imagery in the classification process which adds an element of stand texture, a measure which corresponds to crown size and density, which enables the algorithms to more accurately model tree size. The other is that, based on the FIA estimates, the majority of the samples (approximately 75%) fall within 2 tree size classes (5-10" and 10-15"), where it then becomes statistically more likely to be correctly labeled. Even so, there is very good delineation between these two classes with individual User accuracies of 82% and 92% respectively.

Table 8. Tree Size Class Error Matrix

Flathead National Forest VMap V12; Tree Size Class						
	AA Data					
VMap Class	0-5" DBH	5-10" DBH	10-15" DBH	15" + DBH	Grand Total	User Accuracy
0-5" DBH	14	1	2		17	82%
5-10" DBH	1	81	12	5	99	82%
10-15" DBH		5	92	3	100	92%
15" + DBH			2	42	44	95%
Grand Total	15	87	108	50	260	Overall Accuracy
Producer Accuracy	93%	93%	85%	84%		88%

2f. Extrapolation to the Project Level

The VMap database is constructed and delivered as a mid-level data product. As such, the accuracy assessment is directly applicable to and supports all mid-level applications and analysis conducted with the data. It is recognized, however, that there is a need to use the data at the base level to support Project analysis and the question immediately arises over the applicability of the accuracy assessment to support such use of the data. The answer is that, yes, generally these numbers will support Project analysis but it depends also on the questions that need to be answered and the specific concerns of the Project (Czaplewski and Patterson, 2003).

As a general rule of thumb the accuracy assessment for Lifeform, Tree Canopy Cover, and Tree Size Class will be directly transferrable to the Project specific dataset. Due to the nature of the image classification process each class has an equal probability of classification and is not limited by any spatial constraints or spatial auto-correlation of the remotely sensed data. Also, the manner in which the accuracy assessment was conducted provided a spatially balanced sample against which the classification was tested. Combined, this gives the User confidence that the these classes were tested sufficiently to support analysis conducted using these classes. Keep in mind, however, that this does not preclude the qualitative evaluation of the dataset by the User to ensure that there is not some anomaly within the database that would negatively impact its use for specific Project analysis.

The same could be said for the more prevalent tree dominance types that were mapped across the Flathead National Forest, namely MX-ABLA, MX-PICO, and MX-PSME, as there were sufficient samples against which these types were evaluated. This statement could also apply to MX-LAOC and MX-PIEN but those classes are bumping up against the minimum N for statistical significance (generally accepted to be 30) ((Czaplewski and Patterson, 2003) at 24 and 28 respectively. This would depend entirely upon the Project specific needs of the analysis. For example, if Lynx habitat (which is assumed to correlate highly with MX-PIEN) were of specific concern then it might behoove the User to bolster their confidence in the classification by conducting a more thorough assessment of DOM_MID_40 accuracy throughout the project area. The same can be said of those classes that clearly did not have sufficient sample size for assessment: MX-PIPO, MX-PIAL, MX-POPUL, and HMIX. Analysis that is specifically concerned with these types should have a quantitative accuracy assessment conducted to support the decisions made using those data.

For information specific to Project level use of the VMap database please see the document *Mid-level and Base-level Databases of the R1 Existing Vegetation Mapping Program (VMap)* (Barber, et. al. 2012)

3. Literature Cited

Barber, J. A., Ahl, R.S., and Brown, S.R. (2012), Mid-level and Base-Level Databases of the R1 Existing Vegetation Mapping Program (VMap). USDA Forest Service, Northern Region, Missoula, MT. Numbered Report 12-07-GEO.

Congalton, R.G. (1991), A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment* 37:35-46

Czaplewski, R.L. and Patterson, P.L. (2003), Classification Accuracy for stratification ith remotely sensed data. Forest Science. 49(3): 402-408

Stehman, S.V. and Czaplewski, R.L. (1998), Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sensing of Environment* 64:331-344